LA-UR -87-1789

LA-UR--87-1789

DE87 010137

TITLE SOURCES OF DATA IN THE GENBANK DATABASE

AUTHOR(S) Christian Burks

SUBMITTED TO Proceedings of the First CODATA Workshop on Nucleic Acid and Protein Sequencing Data
Gaithersburg, MD, May 3-6, 1987

DISCLAIMER

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

SOURCES OF DATA IN THE GENBANK DATABASE

Christian Burks

Theoretical Biology and Biophysics Group
T-10, MS K710
Los Alamos National Laboratory
Los Alamos, NM 87545
U.S.A.

(telephone, 505-667-6683; network, cb@lanl.gov)

ABSTRACT

We have characterized the citations in GenBank* with respect to form of publication, and further discuss the potential spectrum of quality control applicable to the various types of unpublished data.

INTRODUCTION

As the number of nucleotides being sequenced has increased, so has the number of journal pages devoted to displaying sequences. This situation has reached the point where many journals are reducing or eliminating pages devoted to displaying sequence data, which raises the question of what the direct sources of sequence data (and quality control of sequence data) are now and will be in the future.

Although data in the GenBank nucleotide sequence database (Burks et al., 1985) have been drawn primarily from articles in refereed journals, the database occasionally includes data from other sources, including unpublished data submitted directly to the data bank by the group determining the sequence. In light of the perception that the literature has been (and will continue to be) overburdened with sequence data, we have sought to quantify the patterns of reference in the GenBank database. (Note that GenBank has shared the task of collecting data from the published literature with the EMBL Data Library (Hamm & Cameron, 1986), so that the coverage of the literature in either database reflects the joint efforts of both.)

We have counted the total number of citations for data in the GenBank database and then broken these down according to form of citation (published or otherwise). We have also determined which journals have reported the

*"GenBank" is a registered trademark of the U.S. Dep't of Health and Human Services.

majority of sequence data up to now. Finally, we elaborate on the notion of "unpublished" data with respect to the data banks' ability to subject submitted data to routine quality control checks.

OVERALL SOURCES OF DATA

The JOURNAL line type of Release 49.0 (April 1987) of the GenBank database (Burks et al., 1985) was scanned to determine the total number and source of citations in the database. The initial scan resulted in about 14430 citations being found. Of course, many of these represent duplicate citations (i.e., a single paper reporting more than one sequence and therefore being cited more than once in the database); after removing these and a small number of citations to articles in press, about 8205 unique citations remained. (The scanning and elimination "algorithms" used were not exact, so the numbers described here have been rounded off to the nearest multiple of 5, and may be inaccurate by +/- 5-10 citations.)

(*** Table I ***)

The distribution of these articles among Ph.D. theses, articles in books, articles in journals, and unpublished sources is presented in Table I. As is clearly seen in Table I, the overwhelming majority of the data in the database has come from articles published in refereed journals. On the other hand, a significant number of sequences are cited to unpublished sources. These unpublished data are, upon examination of the individual GenBank entries involved, most often linked to related published data, though the type of link varies considerably (see the Discussion below).

DISTRIBUTION OF DATA AMONG JOURNALS

How many journals have reported sequence data, and is the distribution among these journals relatively even? In the scan described above, 77 different journals were cited. (Though the total number of journals that have ever included a report of a sequence is presumably much larger, it is unlikely that any single journal not already cited in GenBank has reported significant numbers of sequences over the past five years.) All journals that have published articles corresponding to more than 100 unique citations in the database are presented in Table II. The distribution of sequence data among these journals is quite skewed (see Table II), with about half the data coming from only four journals.

(*** Table II ***)

Note that the relative percentages attributed to any single journal are likely to fluctuate from release to release of GenBank, depending not only on GenBank's own direct data entry but also on whether or not and how much of the most recent releases from the EMBL Data Library (Hamm & Cameron, 1986) have been incorporated.

THE SPECTRUM OF UNPUBLISHED DATA

At the heart of the question of whether or not the data banks should include data from unpublished sources is the assumption that there is a vast difference between published and unpublished data. Though this assumption is in part justified, it should be pointed out that the distinctions between these two sources of data are somewhat blurred in current practice; furthermore, there are different types -- with respect to amenability to quality control checks -- of unpublished data.

Note first that there are at least two viewpoints from which one can judge the relative merits of published and unpublished data: the sociological viewpoint (e.g., Will a resumé with a list of citations to otherwise unpublished data bank entries be as impressive as one citing published journal articles?) and the scientific viewpoint (e.g., Will unpublished data in the data banks be as reliable as published data?). Though these two viewpoints are interrelated, we will restrict our discussion of published and unpublished data to the scientific viewpoint.

Our first point is that the distinction, based on concerns about quality control, between published and unpublished data is not always clear-cut. It has for some time been the case that sequencing gels are not available to a referee reviewing a paper reporting new sequence data. In addition, most sequences being reported are unavailable to the referee in other than hardcopy form (e.g., a copy of the figure that presents the sequence), and are thus not amenable to many of the computer-based routine checks for internal consistency that one normally applies to a sequence (e.g., Does it contain nonsense characters? Is it as long as the text says it is? Are internal stop codons absent from the indicated protein-coding regions?). Thus, published sequences -- as explicitly presented results -- are often not reviewed nearly as carefully as either the experimental protocols or the implications of the sequence data for our understanding of a sub-domain of molecular biology. On the other hand, data banks always end up with a computer-readable version of a sequence (whether submitted directly by the group that determined the sequence or entered by the data bank staff), and thus are able to apply routine checks to every sequence they maintain. The data bank staffs have more than occasionally discovered that the sequence data presented in a published figure bear little resemblance to the sequence being described in the text of that article.

Our second point is that there are several types of unpublished data (and the fact that there are several types should be recognized when evaluating the appropriateness of entering unpublished data into the databases). Consider the four following categories:

Data unrelated to previously published or unpublished data in the GenBank and related databases. Though there are many quality control checks of internal consistency that these data can be subjected to, there is little that can be done to check their links to previously determined sequences. Note that there are very few (if any) unpublished data of this type in the current GenBank database.

Data related to previously published or unpublished data present in the database. There are many examples of this type of data in GenBank; the most frequent are unpublished revisions (by the same group of experimentalists) of older, published data. In addition to internal consistency checks, one can compare the results of aligning the two sets of overlapping information with what the contributors of the unpublished data have claimed; annotation (e.g., the name of the organism, the name of the products, etc.) can also be unified with respect to the standard vocabularies in the rest of the database.

Data described in but not explicitly presented in a published article. The three longest sequences currently in the GenBank database -- all over 100000 nucleotides long -- fall into this category. Though the generation and interpretation of the sequences for the Epstein-Barr virus genome (Baer et al., 1984), the tobacco chloroplast genome (Shinozaki et al., 1986), and the liverwort choloroplast genome (Ohyama et al., 1986) were discussed at great length in journal articles, the articles did not present the actual sequence data. Since, in articles like these, the experimental protocols for and the conclusions drawn from the sequencing have presumably been subjected to journal-mediated review, the sequence data would be expected to have been

double-checked and revised in response to questions and objections that arise during the review process.

Data described in but only partially explicitly presented in a published article. There are also many examples in the GenBank database where authors, when submitting sequence data that appeared in a published article, include more data than was actually published. For example, Nathans and Hogness (1984) published a sequence for the human opsin gene; the sequence presented in the original paper contained about 3000 nucleotides corresponding to the gene's five exons, 5' flank, and 3' flank. When the authors submitted the sequence to GenBank, they included an additional 4000 nucleotides corresponding to the gene's four introns. As with the unpublished data described in the previous paragraph, the quality of this type of unpublished data most likely benefits from the link to a published report. One might assume that there is ar additional benefit arising from the fact that part of the sequence data actually appeared in the paper and was thus directly available to the reviewers; insights into improving the published sequence data would be expected to have migrated into the unpublished (but contiguous) data as well.

DISCUSSION

Though it is clear that thus far the GenBank database has relied predominantly on articles published in refereed journals for the source of nucleotide sequence data, there are a significant number of citations of unpublished sources in the database (note that if one limits a search for citations to data appearing in the last year or two, the relative amount of unpublished nucleotide sequence data in the GenBank database is even greater). The few journals that have carried the majority of the data are becoming increasingly reluctant about providing this role. Though this reluctance will

in part have the effect of forcing sequence data into other journals, it appears more likely that much of the data will appear primarily in the nucleotide sequence data banks. The most important question to address in anticipating -- and perhaps facilitating -- this transition is how one can maintain high standards of quality in sequence data that have not been subjected to journal-mediated peer review. In fact, the data banks are already applying quality control checks within a limited number of contexts that are probably not being applied during journal-mediated review; it is incumbent on the data banks to continue to foster and develop links with molecular biology community that will allow the additional benefits of the type of review usually associated with a published article to become available to database entries.

ACKNOWLEDGEMENTS

We are indebted to D. Nelson, M. Smith, and B. Foley for many past discussions of patterns of citation in the GenBank database. We are also grateful for proofreading by J. Moody. GenBank is funded by a contract (N01-GM-2-2127) with NIGMS which currently includes support from the following co-sponsors: NCI, NIAID, NIADDKD, DRR, NLM, DOE, NSF, USDA, and DOD. This work was also supported under the auspices of the U.S. Dept of Energy.

REFERENCES

Baer,R., Bankier,A.T., Biggin,M.D., Deininger,P.L., Farrell,P.J., Gibson,T.J., Hatfull,G., Hudson,G.S., Satchwell,S.C., Seguin,C., Tuffnell,P.S. and Barrell,B.G. (1984) DNA sequence and expression of the B95-8 Epstein-Barr virus genome. Nature 310, 207-211.

Burks,C., Fickett,J.W., Goad,W.B., Kanehisa,M., Lewitter,F.I., Rindone,W.P., Swindell,C.D., Tung,C.-S. and Bilofsky,H.S. (1985) The GenBank nucleic acid sequence database. Comp. Applic. Biosci. 1, 225-233.

Hamm,G. and Cameron,G. (1986) The EMBL Data Library. Nucl. Acids Res. 14, 5-9.

Nathans,J. and Hogness,D.S. (1984) Isolation and nucleotide sequence of the gene encoding human rhodopsin. Proc. Nat. Acad. Sci. USA 81, 4851-4855.

Ohyama,K., Fukuzawa,H., Kohchi,T., Shirai,H., Sano,T., Sano,S., Umesono,K., Shiki,Y., Takeuchi,M., Chang,Z., Aota,S.-I., Inokuchi,H. and Ozeki,H. (1986) Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. Nature 322, 572-574.

Shinozaki,K., Ohme,M., Tanaka,M., Wakasugi,T., Hayashida,N., Matsubayashi,T., Zaita,N., Chunwongse,J., Obokata,J., Yamaguchi-Shinozaki,K., Ohto,C., Torazawa,K., Meng,B.Y., Sugita,M., Deno,H., Kamogashira,T., Yamada,K., Kusuda,J., Takaiwa,F., Kato,A., Tohdoh,N., Shimada,H. and Sugiura,M. (1986) The complete nucleotide sequence of tobacco chloroplast genome: Its gene organization and expression. EMBO. J. 5, 2043-2049.

TABLES

Table I: Sources of Data in GenBank.

Source	Number of Unique Citations
Ph.D. Thesis	5
Book	15
Unpublished Source[1]	≥265
Journal Article	7920
Total	≥8205

[1]The number given for unpublished citations is probably a slight underestimate because the method -- elimination of pattern-identical citations -- used to trim duplicates most likely removed citations that, though unique, were not distinguishable from other unpublished citations.

Table II: Distribution of Data Among Cited Journals.

Journal[1]	Number of Citations (percentage of total)	Running Total Percentage
Proc. Nat. Acad. Sci. USA	16.7 %	16.7 %
Nucl. Acids Res.	15.0	31.7
J. Biol. Chem.	10.5	42.2
Cell	9.5	51.7
Nature	5.9	57.4
Gene	5.4	62.8
J. Virol.	4.8	67.6
Mol. Cell. Biol.	4.1	71.7
J. Mol. Biol.	3.6	75.5
EMBO J.	3.4	78.9
J. Bacteriol.	3.1	82.0
Science	2.3	84.3
Virology	2.2	86.5
FEBS Lett.	1.6	88.1
Mol. Gen. Genet.	1.6	89.7
Eur. J. Biochem.	1.5	91.2
Biochemistry-USA	1.4	92.6
Biochem. Biophys. Res. Commun.	1.4	94.0

[1]The 18 journals cited at least 100 times in the GenBank database (determined as described in the text). The three groupings of journals roughly demarcate the 50th, 75th, and 94th percentiles.